DOCUMENT RESUME

ED 439 144 TM 030 679

AUTHOR Lee, Guemin
TITLE Estimating Conditional Standard Errors of Measurement for Tests Composed of Testlets.
PUB DATE 1998-12-04
NOTE 46p.; Paper presented at the Annual Meeting of the Iowa Educational Research and Evaluation Association (Ames, IA, December 3-4, 1998).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Definitions; *Error of Measurement; Estimation (Mathematics); *Reliability; Statistical Bias; Test Items
IDENTIFIERS *Testlets

ABSTRACT

The primary purpose of this study was to investigate the appropriateness and implication of incorporating a testlet definition into the estimation of the conditional standard error of measurement (SEM) for tests composed of testlets. The five conditional SEM estimation methods used in this study were classified into two categories: item-based and testlet-based methods. When individual items are used as the fundamental measurement unit, the assumptions required by measurement modeling for tests composed of testlets are violated. Therefore, item-based estimation methods might introduce some magnitude of bias in the estimates of conditional SEMs for tests composed of testlets. In general, the item-based methods provide lower estimates of the conditional SEM along the score scale than do the testlet-based methods. This result is consistent with the previous findings that the reliability of test scores composed of testlets would be overestimated by item-based reliability estimation methods. (Contains 8 tables, 12 figures, and 23 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 439 144

Estimating Conditional Standard Errors of Measurement for Tests Composed of Testlets

Guemin Lee

Paper Presented at the 1998 Annual Meeting of the Iowa Educational Research and Evaluation Association
Ames, IA
December 4, 1998

TM030679

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Guemin Lee
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

BEST COPY AVAILABLE

Abstract

The primary purpose of this study was to investigate the appropriateness and implication of incorporating a testlet definition into the estimation of the conditional standard error of measurement (SEM) for tests composed of testlets. The five conditional SEM estimation methods used in this study were classified into two categories: item-based and testlet-based methods. When individual items are used as the fundamental measurement unit, the assumptions required by measurement modeling for tests composed of testlets are violated. Therefore, item-based estimation methods might introduce some magnitude of bias in the estimates of conditional SEMs for tests composed of testlets. In general, the item-based methods provide lower estimates of the conditional SEM along the score scale than do the testlet-based methods. This result is consistent with the previous findings that the reliability of test scores composed of testlets would be overestimated by item-based reliability estimation methods.

Estimating Conditional Standard Errors of Measurement for Tests Composed of Testlets

In classical test theory, the standard error of measurement (SEM) is estimated by $\hat{\sigma}_E = S_X\sqrt{1-\hat{\rho}_{XX'}}$, where S_X is the standard deviation of a set of test scores and $\hat{\rho}_{XX'}$ is the reliability estimate for those test scores. This formula, which can be viewed as an average standard error of measurement, provides one estimate for all examinees, regardless of their score level (Qualls-Payne, 1992). However, it is reasonable to expect that the amount of error associated with individuals' scores could vary depending on where their true scores are located on the score scale. Since the first edition of the Test Standards, the American Psychological Association, American Educational Research Association and National Council on Measurement in Education (1954), have recommended that test publishers estimate and report the SEM at several points on the score scale. The current version, Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1985), also included this recommendation in Standard 2.10.

Testlets, as the name implies, could be defined as small tests that are small enough to manipulate but large enough to carry their own context (Wainer & Kiely, 1987; Wainer & Lewis, 1990). Previous studies dealing with test scores obtained from tests composed of testlets have indicated that the conditional independence assumption is likely to be violated, making it difficult to satisfy the unidimensionality assumption required by measurement modeling. That is, when several items in a test are related to a common passage or other common stimulus material, dependence is present among those items, meaning that conditional dependence exists (Sireci, Thissen & Wainer, 1991; Yen, 1993; Wainer, 1995; Wainer & Thissen, 1996; Lee & Frisbie, in press; Lee, Kolen, Frisbie & Ankenmann, 1998; Lee, 1998). Under this circumstance, the use of testlet as the unit of analysis, instead of individual items is recommended to eliminate the influence of the dependence among within-passage items (Thissen, Steinberg & Mooney, 1989).

Because test scores from tests composed of testlets would likely to violate the assumptions for measurement modeling, applying unidimensional measurement models based on dichotomously-scored items to estimating conditional SEMs for tests composed of testlets might be inappropriate. Because there is little evidence in the literature about how the violation of assumptions affects estimates of the conditional SEM, it is not clear how serious the degree of distortion of the conditional SEM estimates would be. Testlet-based estimation methods might be considered as alternatives to the item-based estimation methods in estimating the conditional SEM for tests composed of testlets. The purpose of this study was to investigate the appropriateness and implication of adopting a testlet definition to estimating the conditional SEM for tests composed of testlets.

The objectives of this study were to:

1. Assess the dimensionality and conditional dependence of tests composed of testlets to determine the appropriateness of the measurement models that use items as the measurement unit in the context of estimating conditional SEM for these tests.

2. Determine the appropriateness of adopting the testlet concept in estimating the conditional SEM for tests composed of testlets by comparing the differences in estimates between item-based and testlet-based estimation methods using the results from randomly-formed testlets as a criterion.

3. Investigate the relationship between the degree of violation of the assumptions required by measurement modeling and the degree of bias in estimates of the conditional SEM when item-based estimation methods are used instead of testlet-based estimation methods.

Methods of Estimating Conditional SEM

A number of methods have been developed to estimate the conditional SEM. The earliest investigators about the conditional SEM were probably Mollenkopf (1949) and Thorndike (1951). Lord (1955, 1957) developed the best-known conditional SEM estimation formula using binomial error theory. Feldt (1984) provided another estimation method using a compound binomial error model,

which presumes that parallel forms involve stratified random samples of items. An item response theory (IRT) approach to estimating the conditional SEM was provided by Lord (1980), and recently a generalizability theory (G-theory) approach was presented by Brennan (1996). These methods can be thought of as the fundamental frameworks for estimating conditional SEMs, and several variations of these basic frameworks may be possible. A comprehensive review of most of these and related methods is summarized in Feldt & Brennan (1989) and Feldt & Qualls (1996).

Despite all of the works referenced above, the issues related to estimating the conditional SEM for tests composed of testlets have not been addressed. (Brennan (1996) investigated this issue under a generalizability theory framework, however, he did not mention the testlet concept explicitly.) For this study, the estimation methods for the conditional SEM were classified into two categories: item-based and testlet-based methods. IRT and G-theory approaches were considered for estimating the conditional SEM for each item-based and testlet-based method. Because Lord's binomial error model (1955, 1957) and Feldt's compound binomial error model (1984) are special cases of the G-theory approach for estimating the conditional SEM (Brennan, 1996), the IRT and G-theory approaches together include almost all basic formulas discussed above, except variations from Thorndike's (1951) and Mollenkopf's (1949) methods.

A G-theory approach with a pxI design, where p represents persons, the object of measurement, and I represents the item facet, and a dichotomous IRT approach were considered as the item-based estimation methods. A G-theory approach with a px(I:H) design, where p represents persons, H represents the passage facet, and I represents the item facet within a passage, and polytomous IRT approaches for estimating conditional SEM using both Samejima's (1969) graded response model and Bock's (1972) nominal model were used as the testlet-based estimation methods.

Conditional Independence and Unidimensionality

The testlet concept is profoundly related to the conditional independence assumption for measurement modeling. Three conditional dependence indexes were used in this study to investigate the degree of conditional dependence of test scores composed of testlets. First, Yen's (1984) Q_3 statistic can be understood as a correlation of the residuals of an item pair over examinees. Even though the Q_3 statistic is a correlation between residuals of an item pair based on IRT models (therefore, zero correlation might be expected for a conditionally independent item pair), Q_3 has a tendency to be slightly negative in the null case (Yen, 1984, 1993; Chen & Thissen, 1997). Yen (1993) demonstrated that the expected value of the Q_3 statistic, when conditional independence is true, is approximately $-1/(n-1)$, where n is the number of test items. These values can be used as a criterion for comparing the overall level of conditional dependence of within- and between-passage item pairs.

Second, the G^2 statistic is based on the idea of a contingency table. For each pair of items with binary responses, two types of two-way (or four-fold) contingency tables can be constructed–one for observed frequencies and one for expected frequencies. Under the normality assumption about the theta distribution and known item parameters, the G^2 statistic is distributed as χ^2 with one degree of freedom when the number of examinees is large.

Third, the standardized ϕ coefficient difference is distributed asymptotically as the standard normal distribution, N(0,1), if item parameters are known and the theta distribution is normal. This index has an advantage over the G^2 in that it can indicate the direction of association. That is, a positive value indicates greater dependence of the observed frequencies than the IRT model predicts. However, a disadvantage is that this index cannot be defined when the observed frequency in some of the cells is zero (Chen & Thissen, 1997).

A meaningful definition of dimensionality should be based on the principle of conditional independence. The dimensionality of a set of test items can be defined as the number of traits or

latent variables needed to satisfy the assumption of conditional independence (Hambleton & Swaminathan, 1985; Hambleton, 1989; Hambleton, Swaminathan & Rogers, 1991). Considering this inseparable relationship between the conditional independence and the dimensionality of a test or set of items, investigations about the unidimensionality of tests or items could also be considered as an indirect check for the conditional independence assumption. Three methods were used for assessing the unidimensionality of tests composed of testlets in this study.

First, the principal component analysis approach was selected for this study because of its historical tradition and its wide-spread use. Eigenvalues from the inter-item correlation matrix are plotted and then a somewhat subjective judgment needs to be made to assess the dimensionality of the test items. Tetrachoric inter-item correlations are frequently recommended for principal component analysis because other measures of association, like phi correlations, may detect a second spurious factor, which could be identified as a difficulty factor (Carroll, 1945; Hattie, 1985; Hambleton & Swaminathan, 1985; Hambleton, 1989; Roznowski, Tucker & Humphreys, 1991).

Factor analysis is different from principal component analysis in that it estimates a uniqueness for each item given a specified number of factors (Hattie, 1985). When the maximum likelihood estimation method is used, assuming normality, the hypothesis about the number of factors can be tested, assuming a reasonable sample size, using a chi-square test. However, if the data are from binary responses, it would be difficult to meet a multivariate normality assumption and statistical tests can not be used in this situation. McDonald (1982) has suggested that the residual covariance matrix supplies a reasonable basis for judging the extent of the misfit of the one factor model to the data, even though it does not provide the basis for a statistical judgment. This idea was incorporated by Lee, Kolen, Frisbie & Ankenmann (1998): they computed the root mean squares (RMS) of the off-diagonal residuals under each specified number of factors and mainly compared the difference between the RMSs of the one factor model and the two factor model with the difference between the RMSs of the two factor model and the three factor model. They found that the results from

this method are consistent with the results from the principal component analysis, but that this method provided more interpretable results.

Stout's definition of essential dimensionality is based on his definition of essential independence, which was described in a previous part of this section under "conditional independence". That is, Stout's essential dimensionality is the minimum number of dimensions necessary for satisfying the assumption of essential independence (Stout, 1987, 1990; Nandakumar & Stout, 1993). The DIMTEST (Stout, Douglas, Junker & Roussos, 1993) computer application program can be used for testing Stout's essential unidimensionality.

Method

Data Sources

The real data for this study were taken from the 1995 Iowa Tests of Basic Skills (ITBS) Form M to Form K equating study. In this study, certain tests of Form K were used and data from students in grades 4, 7 and 8 were used because the test structures used in these three grades are representative of those in grades 3-8. The Reading Comprehension and Maps and Diagrams tests for grades 4 and 7 and the Vocabulary test from grade 8 were used in this study. The Vocabulary test was included because it may be the most unidimensional test in the ITBS test battery. There are 43 items in the Vocabulary test for grade 8 (Hoover, Hieronymus, Frisbie & Dunbar, 1994). The sample size and the general characteristics of each test are presented in Table 1.

Insert Table 1 About Here

A unidimensional simulated data set was created to have the same structure as the Vocabulary test. The simulated response data were generated by following the procedures used by Yen (1984), assuming item parameter estimates of the grade 8 Vocabulary test from the 1992 ITBS national standardization sample as the true item parameters. Though the Vocabulary and simulated

data sets do not have naturally-formed testlets, seven testlets were randomly constructed for the purpose of comparison with tests composed of naturally-formed testlets.

Analyses

The computer program IRT_LD (Chen & Thissen, 1997) was used compute conditional dependence measures for each data set: Yen's Q_3 statistic (Yen, 1984), the likelihood ratio G^2 (Chen & Thissen, 1997), and the standardized ϕ coefficient difference (Chen & Thissen, 1997). To investigate the nature of the conditional dependence measures of within- and between-passage items, distributional characteristics (e.g., mean, standard deviation, skewness, and kurtosis) of the pair of conditional dependence measures (one for within-passage and one for between-passage) were compared. The percentages of the hypothesis rejections (conditional independence hypothesis) for within-passage and between-passage item pairs were compared for each measure.

To investigate the unidimensionality of tests composed of testlets, the principal component analysis, exploratory factor analyses, and Stout's (1987, 1990) essential unidimensionality test were completed. For principal component analyses, tetrachoric correlations were computed first by the PRELIS2 computer program (Joreskog & Sorbom, 1993). Then, after doing the principal component analyses, a scree plot was used to display the results. From sets of factor analyses, the root mean squares (RMS) of the off-diagonal residuals under each specified number of factors were compared. Stout's essential unidimensionality test was conducted by the DIMTEST (Stout, Douglas, Junker & Roussos, 1993) computer application program.

The item-based and testlet-based conditional SEM estimation methods were applied to each data set. For the G-theory approach, a computer application program (Brennan, 1996) was used to estimate the conditional SEM for each pxI or px(I:H) design. For the IRT methods, the BILOG (Mislevy & Bock, 1990) and MULTILOG (Thissen, 1991) computer programs were used for estimating item parameters and ability parameters. The number-correct raw score distribution, given theta, was formulated (Lord & Wingersky, 1984; Hanson, 1994; Wang, Kolen & Harris, 1996)

and the conditional SEM was estimated by a FORTRAN 90 application program written for this purpose.

Data from the Vocabulary test and the simulated data set served as criteria for interpreting the difference between item-based methods and testlet-based methods for tests composed of stimulus-based testlets (Reading Comprehension and Maps and Diagrams tests). The conditional dependence measures of each test were interpreted in connection with the magnitude of the difference between the conditional SEM estimates from the item-based methods and testlet-based methods.

Results

Conditional Independence Assumption Check

Yen's Q_3 statistic was used here as a measure of conditional dependence. If there are n items in a test, $n(n\text{-}1)/2$ Q_3 statistics can be computed. In a similar way, for k_h items in the hth passage, there are $k_h(k_h-1)/2$ Q_3 statistics. Two types of Q_3 statistics were distinguished in this study for each test: one is the within-passage Q_3 statistics (No. of $Q_3 = \sum_{h=1}^{H} k_h(k_h-1)/2$), and the other is the between-passage Q_3 statistics (No. of $Q_3 = n(n\text{-}1)/2 - \sum_{h=1}^{H} k_h(k_h-1)/2$). The distributional statistics for within-passage and between-passage Q_3 conditional dependence measures are shown in Table 2.

Insert Table 2 About Here

The averages of the Q_3 statistics from within- and between-passage item pairs would be similar to the expected values of the Q_3 measures if the conditional independence assumption holds. Table 2 shows that the averages of between-passage Q_3 statistics for Reading and Maps tests for grades 4 and 7 have values similar to the expected values of the Q_3 statistics, implying that item pairs between passages are conditionally independent. In contrast, the averages of within-passage Q_3 statistics for these tests have more positive values compared to the expected values of Q_3. This suggests that the conditional independence assumption is violated. For the Vocabulary test and

simulated data set, because testlets were randomly constructed, averages of within- and between-passage Q_3 statistics are both similar to the expected values of Q_3, as would be anticipated. In comparing the difference between the observed mean and the expected mean of the Q_3 values with the standard deviation of the observed Q_3 statistics, in cases where conditional dependence was identified, the magnitude of the difference seems to be about one standard deviation. On the other hand, where conditional dependence was not identified, the magnitude of the difference is much less than one standard deviation and close to zero.

The likelihood ratio G^2 statistic is distributed as χ^2 with one degree of freedom for a large number of examinees under a normality assumption about the theta distribution and with known item parameters. The main reason for including this statistic is to conduct statistical tests about the conditional independence hypothesis for item pairs. If the distributional assumption about the G^2 statistic is true, it would be reasonable to anticipate the expected value of one and, 5% and 1% rejection rates when 3.84 and 6.63 are used as critical values for the chi-square statistical tests.

According to Table 3, the averages of the G^2 statistics of between-passage item pairs have values similar to one, the expected value of the χ^2 distribution with one degree of freedom, for tests composed of testlets, except the grade 7 Reading test. In contrast, the averages of the within-passage G^2 statistics for both Reading and Maps tests of grades 4 and 7 have values greater than one. Also, the rejection rates for the hypothesis of conditional independence of between-passage item pairs are around 5%, but the rejection rates of within-passage item pairs are over 20% and up to about 40% when 3.84 was used as a critical value. With a critical value of 6.63, about 1% rejection rates are found for the between-passage item pairs, but over 10% rejection rates, and up to about 20%, are observed for the within-passage item pairs for tests composed of testlets.

Insert Table 3 About Here

For the grade 8 Vocabulary test and the simulated data set, similar descriptive statistics about the G^2 statistics were obtained for both between- and within-passage item pairs. Although much higher rejection rates about the conditional independence hypothesis were observed in within-passage item pairs compared to the expected rejection rate for four tests composed of testlets, but the rejection rates of within-passage item pairs for the Vocabulary and simulated data sets were similar to the expected rejection rate.

The standardized ϕ coefficient difference is expected to be distributed as standard normal. This index has an advantage over the G^2 statistic because it has a sign to indicate the direction of association. That is, a positive value of this index represents greater dependence of the observed frequencies than the IRT model predicts, and a negative value represents the opposite case (Chen & Thissen, 1997). In this study, 1.96 and -1.96 were used as the upper and lower critical values. Therefore, a 2.5% rejection rate about the conditional independence hypothesis within each tail of the distribution can be expected in the null case. The rejection rates about the conditional independence hypothesis using the standardized ϕ coefficient difference are presented in Table 4.

Insert Table 4 About Here

The hypothesis rejection rates of between-passage item pairs are around 2.5% for all six data sets, even though some fluctuations are observed in some tests. However, these fluctuations seem to be negligible compared to the rejection rates of within-passage item pairs for the Reading and Maps tests. That is, the hypothesis rejection rates of within-passage item pairs for these four tests are around 20%-30% with the upper side critical value of 1.96. These rejection rates are much greater than the expected rejection rate of 2.5% when the null hypothesis is true.

One important finding can be observed in this table. That is, when using the critical value of -1.96, the hypothesis rejection rates of within-passage item pairs for tests composed of testlets are similar to the rejection rates of between-passage item pairs. This means that the rejection of

conditional independence hypothesis of within-passage item pairs is mainly due to positive association among items within a particular passage. In other words, for a group of items sharing the same stimulus material, it would be reasonable to expect a positive association among those items, which would lead to the rejection of the conditional independence hypothesis among those items.

Unidimensionality Assumption Check

Table 5 provides the first ten eigenvalues from tetrachoric correlation matrices based on individual items from the six data sets. These indicate that more than one factor would be required for explaining the data from the Reading and Maps tests for grades 4 and 7. However, for both the Vocabulary test and the simulated data set, one factor appears to be sufficient to explain the data.

Insert Table 5 About Here

To get more information about the dimensionality of each test, the root mean square (RMS) of the off-diagonal residuals under each specified number of factors was computed, as shown in Table 6. The difference between the RMSs of the one factor model and the two factor model from the Reading and Maps tests for grades 4 and 7 are about two to three times greater than the difference between the RMSs of the two factor model and the three factor model. This means that one factor does not appear to be sufficient to describe the dimensionality of these four tests. For both the Vocabulary test and the simulated data set, the difference between the RMSs of the one factor model and the two factor model is similar to the difference between the RMSs of the two factor model and the three factor model. Here, one factor seems sufficient to describe dimensionality. The results of several exploratory factor analyses, mainly comparing the RMSs, are consistent with the results from the principal component analyses.

Insert Table 6 About Here

Stout's essential unidimensionality test is somewhat different from conventional approaches for assessing the dimensionality of tests or sets of items. That is, this approach basically relies on a definition of essential unidimensionality. In most IRT applications, the unidimensionality assumption is required, but it cannot be strictly satisfied because there are always other cognitive, personality, and test-taking factors that affect test performance (Hambleton, 1989). Tests often are constructed by including several minor factors in addition to the dominant dimension. Therefore, the important thing for the assumption of unidimensionality to be met is to satisfy the assumption about one dominant component or factor. Based on this perspective, Stout (1987, 1990) relaxed the definition of conditional independence and developed procedures to test the essential unidimensionality, statistically based on this relaxed definition of conditional independence (called essential independence in his paper).

The main purpose of Stout's essential unidimensionality test is to assess the lack of unidimensionality of a test, possibly as a preliminary analysis for using the unidimensional IRT models in various application contexts (Stout, Douglas, Junker & Roussos, 1993). That is, even though more than one dimension might be identified by principal component analysis or exploratory factor analyses, using the dominant factor idea, the identified second dimension could be considered a minor factor. Under these circumstances, unidimensional measurement models might be appropriate to use. Checking this possibility is the main reason to conduct Stout's essential unidimensionality test in this study.

According to Table 7, for the Reading and Maps tests for grades 4 and 7, the essential unidimensionality hypothesis is rejected at 0.05 level of significance (based on a critical value of 1.96 for two-tailed test). That is, more than one dominant dimension would be required to explain data from these tests composed of testlets. In contrast, one dominant factor seems to be sufficient to describe the data for both the grade 8 Vocabulary test and the simulated data set.

Insert Table 7 About Here

On the basis the results discussed so far, it might be suspected that unidimensional measurement models based on dichotomously scored items might be problematic in applications involving tests composed of testlets. That is, the common use of the unidimensional dichotomous measurement models in various application situations could be suspect because of violations of assumptions. To check the possibility of adopting measurement models based on testlet scores, principal component analyses with product-moment correlation matrices among testlet scores were conducted. Eigenvalues are presented in Table 8.

Insert Table 8 About Here

One factor is evident, and the other eigenvalues are negligible. The well-known Kaiser (1970) criterion, retaining eigenvalues greater than unity, has been criticized because of its susceptibility to the overidentification of dimensions (Cliff, 1988). Based on the Kaiser criterion, only one dimension is retained for all forms of all tests. In view of this susceptibility to overidentification, unidimensionality can be supported for the tests used in this study when testlet scores are used as the unit of analysis.

Estimating Conditional SEMs

The estimated conditional SEMs based on using five different estimation methods with the grade 4 Reading test are presented in Figure 1. The horizontal axis represents an observed score scale and the vertical axis represents the estimated conditional SEM. The conditional SEM for a given observed score point was computed by summing up conditional error variances of examinees having the same observed score, averaging the summed error variances, and taking the square root of the average.

Insert Figure 1 About Here

In order to get a basis for interpreting these results, it would be helpful to consider a finding from previous studies related to estimating the reliability of test scores composed of testlets (Sireci, Thissen & Wainer, 1991; Wainer, 1995; Wainer & Thissen, 1996; Lee & Frisbie, 1997). They consistently indicated that the conventional reliability estimation methods based on item scores, like coefficient alpha, overestimate the reliability of test scores composed of testlets. Consequently, it is reasonable to expect the item-based conditional SEM estimation methods would underestimate the conditional SEM of test scores composed of testlets. Five estimation methods used in this study were classified as item-based (G-theory approach with a pxI design and dichotomous IRT approach) or testlet-based (G-theory approach with a px(I:H) design and two polytomous IRT approaches) methods.

According to Figure 1, the estimation method based on the dichotomous three-parameter logistic model [DIRT method] provided the lowest estimates of the conditional SEM, except in the score range from 1 to 7. For scores less than 7, the estimated conditional SEMs were almost the same, making a nearly horizontal line. This might be explained by a guessing effect associated with multiple–choice items. That is, because the three-parameter logistic model was used as a fundamental model for estimating conditional SEM, it would be natural for an examinee having very low ability to get a score of about 7 or 8 [the total number of items * 1/(the number of choices +1) = 38 * 1/(4+1) = 7.6]. As previously mentioned, because the observed score scale was used for the horizontal axis in this study, this method could not differentiate the conditional SEMs for scores less than 7. Samejima's graded response model [GIRT method] and Bock's nominal model [NIRT method] estimation methods provided similar conditional SEM estimates, but the GIRT method provided slightly larger estimates, especially in the lower score range.

The estimates of conditional SEM from a G-theory approach with a pxI design [pxI method] were lower than those from a G-theory approach with a px(I:H) design [px(I:H) method]. The pxI method provided higher estimates than did the DIRT method, but lower estimates compared to the GIRT and NIRT methods. The estimates of conditional SEM from the px(I:H) method were highest in

the middle score range (from 10 to 30), but the NIRT and GIRT methods provided higher conditional SEM in both the lower and higher score ranges on the observed score scale. The px(I:H) method provided the most irregular curve, while the other four methods form smooth curves.

For a more convenient comparison, the differences in conditional SEM estimates between item-based and testlet-based methods were computed by subtracting the conditional SEM estimates of item-based method from those of testlet-based method, as presented in Figure 2. The G-theory approaches and IRT approaches were graphed separately because a direct comparison between G-theory approaches and IRT approaches would require a baseline, which does not exist in this study.

In the top graph of Figure 2, it can be seen that both GIRT and NIRT provided higher estimates of the conditional SEM than the item-based method, DIRT, over the usable score range. The difference was more evident in the high score range. The bottom graph shows that the px(I:H) method provided much higher estimates of conditional SEM than did the pxI method. Because estimates of conditional SEM for the px(I:H) method are very bumpy, the differences in conditional SEM estimates between the pxI and px(I:H) methods also make an irregular curve. However, in general, a kind of concave-downward quadratic line could be imagined, which means that the differences between the two estimation methods are bigger in the middle score range than they were in the extreme lower and higher score ranges.

Insert Figure 2 About Here

Similar trends can be found for the grade 7 Reading test. These are presented in Figure 3. The main difference in the results between the grade 4 and 7 Reading tests was that the discrepancy of the conditional SEM estimates between the pxI and px(I:H) methods was much more evident in grade 7. Another main difference is that the NIRT and GIRT methods provided somewhat different estimates of the conditional SEM in the lower score range (especially in the score points less than 16). These two observations are more evident in Figure 4, which represents the differences in conditional

SEM estimates between the item-based and testlet-based estimation methods. The bottom graph shows that it is still reasonable to represent the data in terms of a concave-downward quadratic curve. However, the magnitude of differences in conditional SEM estimates between the pxI and px(I:H) methods is much greater than those from the grade 4 Reading test. This trend also can be found in the top graph, even though it is less clear compared to the bottom graph.

Insert Figure 3 About Here

Insert Figure 4 About Here

The estimates of conditional SEM from using each estimation method for the grade 4 Maps test are presented in Figure 5, and the differences in conditional SEM estimates between item-based and testlet-based methods are shown in Figure 6.

Insert Figure 5 About Here

Insert Figure 6 About Here

The basic trends are the same as found in the Reading tests for grades 4 and 7. However, there are several important distinctions between this figure and the previous figures. First, in the middle score range (around from 10 to 18) the pxI method provided estimates in the conditional SEM that are not easily differentiated from the estimates of the GIRT and NIRT methods. Second, the differences of conditional SEM estimates between item-based and testlet-based estimation methods are smaller compared to the Reading tests for both grades. Third, the pxI and px(I:H) methods provided a conditional SEM estimate of zero for a perfect score (in this case, a score of 26). (The estimate of zero is possible for G-theory approaches because the application program for estimating the conditional SEM with G-theory approaches (Brennan, 1996) allows such estimates for a perfect or zero total test

score.) Fourth, the similarity between the GIRT and NIRT estimation methods is more evident compared to the Reading tests.

The results of estimating conditional SEM for the grade 7 Maps test are presented in Figure 7. Basically, trends for the grade 4 Maps test appear here also, except for the conditional SEM estimates from the px(I:H) method. The px(I:H) method provided the highest estimates of conditional SEM in the middle score range for the other tests, but in the grade 7 Maps test, it provided conditional SEM estimates similar to those from the GIRT and NIRT estimation methods. The differences in conditional SEM estimates between item-based and testlet-based methods are graphed in Figure 8. Trends in this figure are similar to those shown in Figure A.4, which describes the results from the grade 4 Reading test.

Insert Figure 7 About Here

Insert Figure 8 About Here

The grade 8 Vocabulary test and a simulated data set were included in this study for the purpose of comparison with tests composed of naturally-formed testlets. In each case, dimensionality is controlled by conception and design. The Vocabulary test may be the most unidimensional test in the ITBS test battery, and the simulated data set also can be considered unidimensional because a unidimensional IRT model was used to produce it. For comparing with other tests composed of naturally-formed testlets, seven testlets were randomly constructed and the polytomous IRT models were applied to estimate the conditional SEM. Similar conditional SEM estimates would be expected to be observed for both dichotomous and polytomous IRT estimation methods and for both pxI and px(I:H) methods. The estimated conditional SEM for the grade 8 Vocabulary test and simulated data set are presented in Figures 9 and 10, respectively, and differences in conditional SEM estimates

between item-based and testlet-baed estimation methods are shown in Figures 11 and 12, respectively.

Insert Figure 9 About Here

Insert Figure 10 About Here

Insert Figure 11 About Here

Insert Figure 12 About Here

The DIRT, px(I:H), GIRT, and NIRT methods provide similar estimates of conditional SEMs, even though the curve of the px(I:H) method is less smooth. Two important observations can be made from these results. First, for both the Vocabulary test and the simulated data set, the pxI method provides the highest estimates of conditional SEM in the middle score range (around from 15 to 30). The second observation is that the NIRT and GIRT estimation methods provide higher estimates of the conditional SEM than the other methods in the highest score range (scores over 35).

Discussion

Five main generalizations follow from the findings of this study:

First, when items are used as the fundamental measurement unit, the assumptions required by measurement modeling (conditional independence and unidimensionality) for tests composed of testlets are violated, but those assumptions are satisfied when testlets are used as the measurement unit. Therefore, the unidimensional measurement model based on dichotomously scored items may be inappropriate for estimating conditional SEM for tests composed of testlets. In contrast, the use of a unidimensional measurement model based on testlet scores can be advocated because it satisfies these assumptions.

Second, for the Reading tests, the DIRT method provides the lowest estimates of the conditional SEM compared to the other estimation methods. The px(I:H) method provides higher conditional SEM estimates in the middle score range, but both polytomous IRT estimation methods give higher estimates in the higher score range. The px(I:H) method provides the most irregular curve: the other methods give smooth curves. In general, the item-based methods provide lower estimates of the conditional SEM than do the testlet-based methods. This is consistent with previous findings related to methods of estimating reliability.

Third, for the Maps tests, the basic trends are similar to those found with the Reading tests. However, in the grade 4 Maps test, in the middle score range the pxI method provides estimates of conditional SEM that are not easily differentiated from the estimates from the GIRT and NIRT methods, and the differences in conditional SEM estimates between item-based and testlet-based estimation methods are much smaller. This can be explained in terms of the relationship between the degree of violation of the assumptions and its effect on the estimates of the conditional SEM. That is, because the assumptions for measurement modeling based on dichotomously-scored items are less violated in the grade 4 Maps test compared to other tests, the more similar conditional SEM estimates for the item-based and testlet-based methods found for the grade 4 Maps test might not be so surprising. These results form one piece of evidence to demonstrate how the degree of violation of the assumptions required for measurement modeling affects the estimates of the conditional SEM.

Fourth, for both the Vocabulary test and simulated data set, the pxI method provided the highest estimates of conditional SEM in the middle score range. Previous research suggests one possible explanation for this outcome. The pxI method might consistently overestimate the conditional SEM for tests satisfying the unidimensionality and conditional independence assumptions (Lee, Brennan, & Kolen, 1998). Therefore, it would be reasonable to expect the higher estimates of conditional SEMs using the pxI method compared to other estimation methods in the Vocabulary test and the unidimensional simulated data set. Then, the pxI method would provide the robust estimates

of the conditional SEMs under mild violation of the assumptions, and this method is more robust to the violation of the assuptions compared to the DIRT method.

Fifth, the NIRT and GIRT estimation methods provide higher estimates of the conditional SEM in the higher score range for both the Vocabulary test and the simulated data set. Yen (1993) indicated that if some items within a particular testlet are locally independent or less locally dependent, there would be a loss of information when testlet scores are computed and used as the unit of analysis. So the fact that the GIRT and NIRT methods provide higher estimates of the conditional SEM in the highest score range might be explained in terms of a loss of information. That is, whenever there is some loss of information, a relatively higher conditional SEM should be expected. This is evident from what is known about the relationship between an information function and the conditional SEM. Then, it would be reasonable to anticipate the some loss of information may occur on the estreme score ranges, not on the middle score range.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Brennan, R.L. (1996). *Conditional standard errors of measurement in generalizability theory*. Iowa Testing Programs Occasional Paper No. 40. Iowa City, IA: University of Iowa.

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265-289.

Hoover, H.D., Hieronymus, A.N., Frisbie, D.A., & Dunbar, S.B. (1994) *Iowa Tests of Basic Skills : Interpretive guide for school administrators*. Chicago, IL: The Riverside Publishing Company.

Kolen, M.J., Zeng, L., & Hanson, B.A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement, 33*, 129-140.

Lee, G. (1998). *A comparison of methods of estimating conditional standard errors of measurement for tests composed of testlets*. Unpublished doctoral dissertation, The University of Iowa, Iowa Citye, IA.

Lee, G., & Frisbie, D.A. (in press). Estimating reliability under a generalizability theory model for test scores composed of testlets. *Applied Measurement in Education*.

Lee, G., Kolen, M.J., Frisbie, D.A., & Ankenmann, R.D. (1998, April). *Equating test forms composed of testlets using dichotomous and polytomous IRT models*. Paper presented at the Annual Meeting of National Council on Measurement in Education, San Diego, CA.

Lee, W., Brennan, R.L., & Kolen, M.J. (1998, April). *A comparison of some procedures for estimating conditional scale-score standard errors of measurement.* Paper presented at the Annual Meeting of National Council on Measurement in Education, San Diego, CA.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

Samejima, F. (1969). Estimation of a latent ability using a response pattern of graded scores. *Psychometric Monograph Supplement, 17.*

Sireci, S.G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28,* 237-247.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52,* 589-617.

Stout, W., Douglas, J., Junker, B., & Roussos, L. (1993). *DIMTEST manual.* Urbana-Champaign, IL: University of Illinois.

Thissen, D., Steinberg, L., & Mooney, J.A. (1989). Trace lines for testlets: A use of multiple-categorical models. *Journal of Educational Measurement, 26,* 247-260.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 law school admissions test as an example. *Applied Measurement in Education, 8,* 157-186.

Wainer, H., & Kiely, G.L. (1987). Item clusters and computerized adaptive testing : A case for testlets. *Journal of Educational Measurement, 24,* 185-201.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27,* 1-14.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement : Issues and Practice, 15,* 22-29.

Wang, T., Kolen, M.J., & Harris, D.J. (1996). *Conditional standard errors, reliability and decision consistency of performance levels using polytomous IRT*. Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.

Yen, W.M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8,* 125-145.

Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187-213.

Table 1
Descriptive Statistics for Data Sources Used in This Study

Characteristics	Reading Grade 4	Reading Grade 7	Maps Grade 4	Maps Grade 7	Vocabulary Grade 8	Simulated Data
Sample Size	985	629	914	682	666	1000
No. of Items	38	46	26	30	43	43
No. of Passages	8	7	4	5	7	7
No. of Items per Passage	6,5,3,6, 5,4,3,6	7,7,7,9, 4,7,5	6,6,7,7	6,5,5,7,7	7,6,6,6, 6,6,6	7,6,6,6, 6,6,6
$\bar{X}$	19.7	25.4	14.4	13.8	24.4	26.4
S_X	7.44	9.08	5.39	5.68	8.87	8.47
Skewness	0.118	0.080	-0.023	0.431	0.018	-0.196
Kurtosis	2.139	1.968	2.188	2.327	2.170	2.280

Note : Reading = Reading Comprehension, Maps = Maps and Diagrams

Table 2

Distributional Characteristics of Yen's Q_3 Statistic for Within-Passage and Between-Passage Item Pairs

Test	No. of Q_3	E (Q_3)	Mean	Diff	S.D.	Range
Reading (4)	703	-.027				
Between	626		-.025	.002	.038	-.144 ~ .099
Within	77		.029	.056	.064	-.104 ~ .279
Reading (7)	1035	-.022				
Between	899		-.022	.000	.044	-.151 ~ .106
Within	136		.027	.049	.061	-.115 ~ .245
Maps (4)	325	-.040				
Between	253		-.037	.003	.039	-.177 ~ .051
Within	72		.003	.043	.045	-.145 ~ .101
Maps (7)	435	-.034				
Between	358		-.035	.001	.044	-.175 ~ .077
Within	77		.027	.061	.051	-.081 ~ .158
Vocabulary (8)	903	-.024				
Between	792		-.017	.007	.043	-.144 ~ .106
Within	111		-.021	.003	.040	-.147 ~ .089
Simulation	903	-.024				
Between	792		-.019	.005	.034	-.151 ~ .107
Within	111		-.016	.008	.031	-.088 ~ .052

Note. Reading (4) = grade 4 Reading Comprehension, Reading (7) = grade 7 Reading Comprehension, Maps (4) = grade 4 Maps and Diagrams, Maps (7) = grade 7 Maps and Diagrams, Simulation = simulated data; E (Q_3) = Expected value of Q_3, and Diff = Absolute value of the difference between E (Q_3) and the sample mean.

Table 3
Distributional Characteristics of G^2 Statistics and
Percentage of the G^2 Statistics Greater than Two Critical Values

Test	No. of G^2	Mean	S.D.	Range	% > 3.84	% > 6.63
Reading (4)	703				7.5	3.3
Between	626	1.00	1.37	0.00 ~ 9.58	4.6	1.0
Within	77	5.50	10.78	0.05 ~ 73.70	31.2	22.1
Reading (7)	1035				11.0	3.3
Between	899	1.72	1.38	0.11 ~ 12.08	6.9	1.0
Within	136	4.47	5.33	0.01 ~ 36.43	38.2	18.4
Maps (4)	325				9.2	2.8
Between	253	1.02	1.36	0.01 ~ 9.43	5.5	0.8
Within	72	2.33	3.16	0.09 ~ 17.93	22.2	9.7
Maps (7)	435				7.6	3.0
Between	358	0.98	1.19	0.02 ~ 8.30	3.4	0.8
Within	77	3.04	3.93	0.05 ~ 19.76	27.3	13.0
Vocabulary (8)	903				3.9	0.9
Between	792	1.06	1.24	0.01 ~ 8.05	3.8	1.0
Within	111	0.87	1.03	0.03 ~ 6.72	4.5	0.9
Simulation	903				5.4	1.0
Between	792	1.74	1.28	0.21 ~ 12.41	5.6	1.1
Within	111	1.64	0.98	0.50 ~ 5.62	4.5	0.0

Note. Reading (4) = grade 4 Reading Comprehension, Reading (7) = grade 7 Reading Comprehension, Maps (4) = grade 4 Maps and Diagrams, Maps (7) = grade 7 Maps and Diagrams, Simulation = simulated data; % > 3.84 = percentage of the G^2 statistics greater than 3.84, % > 6.63 = percentage of the G^2 statistics greater than 6.63.

Table 4

Distributional Characteristics of the Standardized ϕ Coefficient Difference (ϕ_{diff}) and Percentage of the ϕ_{diff} Statistics Greater than Two Critical Values

Test	No. of ϕ_{diff}	Mean	S.D.	Skewness	Kurtosis	% > 1.96	% < -1.96
Reading (4)	703					5.5	2.0
Between	626	0.07	0.96	0.18	3.26	2.6	2.1
Within	77	1.58	1.84	1.36	6.86	29.9	1.3
Reading (7)	1035					7.0	0.5
Between	899	0.27	0.95	-0.01	2.89	3.1	0.6
Within	136	1.50	1.29	0.65	4.04	32.4	0.0
Maps (4)	325					5.2	3.1
Between	253	-0.10	0.94	0.04	3.30	1.6	3.6
Within	72	0.86	1.27	-0.17	4.02	18.1	1.4
Maps (7)	435					5.3	2.1
Between	358	-0.17	0.91	-0.05	2.91	0.6	2.5
Within	77	1.23	1.23	0.34	3.28	27.3	0.0
Vocabulary (8)	903					3.0	0.9
Between	792	0.38	0.90	-0.09	2.99	3.0	0.9
Within	111	0.27	0.83	-0.15	3.19	2.7	0.9
Simulation	903					2.9	0.4
Between	792	0.30	0.88	-0.12	2.88	2.9	0.5
Within	111	0.35	0.80	-0.06	2.60	2.7	0.0

Note. Reading (4) = grade 4 Reading Comprehension, Reading (7) = grade 7 Reading Comprehension, Maps (4) = grade 4 Maps and Diagrams, Maps (7) = grade 7 Maps and Diagrams, Voc (8) = grade 8 Vocabulary, Simulation = simulated data; % > 1.96 = percentage of the ϕ_{diff} statistics greater than 1.96, % < -1.96 = percentage of the ϕ_{diff} statistics less than -1.96.

Table 5
First Ten Eigenvalues of Tetrachoric Correlation Matrices Based on Individual Item Scores for Six Tests

Eig	Reading (4)		Reading (7)		Maps (4)		Maps (7)		Voc (8)		Simulation	
Rank	Eig	Diff	Eig	Diff	Eig	Diff	Eig	Diff	Eig	Diff	Eig	Diff
1	10.10	8.16	13.26	10.45	7.87	6.24	6.81	5.05	13.59	11.92	12.87	11.45
2	1.94	0.41	2.81	1.13	1.64	0.27	1.76	0.28	1.67	0.19	1.42	0.09
3	1.54	0.12	1.68	0.22	1.36	0.23	1.48	0.12	1.48	0.07	1.32	0.06
4	1.41	0.11	1.46	0.09	1.13	0.08	1.36	0.05	1.41	0.15	1.27	0.06
5	1.30	0.11	1.36	0.03	1.05	0.02	1.30	0.15	1.26	0.06	1.20	0.06
6	1.19	0.06	1.33	0.05	1.03	0.07	1.16	0.04	1.20	0.01	1.14	0.05
7	1.13	0.03	1.28	0.04	0.96	0.04	1.12	0.07	1.19	0.03	1.09	0.03
8	1.10	0.06	1.24	0.04	0.92	0.02	1.05	0.07	1.16	0.06	1.06	0.00
9	1.04	0.01	1.20	0.06	0.90	0.04	0.98	0.00	1.10	0.03	1.06	0.02
10	1.00	0.02	1.14	0.06	0.86	0.08	0.98	0.04	1.07	0.02	1.03	0.01

Note. Reading (4) = grade 4 Reading Comprehension, Reading (7) = grade 7 Reading Comprehension, Maps (4) = grade 4 Maps and Diagrams, Maps (7) = grade 7 Maps and Diagrams, Voc (8) = grade 8 Vocabulary, Simulation = simulated data, Eig = eigenvalue, and Diff = difference between sonsecutive eigenvalues.

Table 6
Root Mean Squares of Off-Diagonal Residuals for a Specified Number of Factors for Six Tests

No.	Reading (4)		Reading (7)		Maps (4)		Maps (7)		Voc (8)		Simulation	
Fac	RMS	Diff	RMS	Diff	RMS	Diff	RMS	Diff	RMS	Diff	RMS	Diff
1	6.0	1.0	7.3	1.6	6.2	1.4	6.6	1.0	5.7	0.5	4.5	0.3
2	5.0	0.5	5.7	0.5	4.8	0.7	5.6	0.5	5.2	0.3	4.2	0.2
3	4.5	0.4	5.2	0.3	4.1	0.3	5.1	0.5	4.9	0.3	4.0	0.2
4	4.1	0.3	4.9	0.2	3.8	0.4	4.6	0.5	4.6	0.2	3.8	0.2
5	3.8	0.2	4.7	0.3	3.4	0.3	4.1	0.3	4.4	0.3	3.6	0.2
6	3.6	0.3	4.4	0.2	3.1	0.3	3.8	0.3	4.1	0.2	3.4	0.2
7	3.3	0.2	4.2	0.3	2.8	0.2	3.5	0.3	3.9	0.3	3.2	0.1
8	3.1	0.2	3.9	0.2	2.6	0.3	3.2	0.2	3.6	0.2	3.1	0.2
9	2.9	0.3	3.7	0.2	2.3	0.2	3.0	0.3	3.4	0.2	2.9	0.2
10	2.6		3.5		2.1		2.7		3.2		2.7	

Note. Reading (4) = grade 4 Reading Comprehension, Reading (7) = grade 7 Reading Comprehension, Maps (4) = grade 4 Maps and Diagrams, Maps (7) = grade 7 Maps and Diagrams, Voc (8) = grade 8 Vocabulary, Simulation = simulated data, No. Fac = number of factors, RMS = root mean square of off-diagonal residuals, and Diff = difference of RMSs for consecutive numbers. The scales of the RMS and difference are changed by multiplying all entries by 100 and then rounding to one decimal place.

Table 7
Results of Stout's Essential Unidimensionality Test for Six Tests

Test	No. of Items	No. of Examinees	T Statistics	
			T-Value	Probability
Reading for grade 4	38	985	4.72	.000
Reading for grade 7	46	629	3.75	.000
Maps for grade 4	26	914	2.00	.022
Maps for grade 7	30	682	2.24	.012
Vocabulary for grade 8	43	666	0.80	.211
Simulation Data	43	1000	0.74	.230

Note. Reading = Reading Comprehension, Maps = Maps and Diagrams and T-value are referred to a normal distribution to determine statistical significance.

Table 8
Eigenvalues of Product-Moment Correlation Matrices Based on Passage Scores for Six Tests

Eig	Reading (4)		Reading (7)		Maps (4)		Maps (7)		Voc (8)		Simulation	
Rank	Eig	Diff	Eig	Diff	Eig	Diff	Eig	Diff	Eig	Diff	Eig	Diff
1	3.63	2.77	3.59	2.66	2.49	1.93	2.46	1.73	4.44	3.95	4.26	3.68
2	0.86	0.07	0.94	0.23	0.56	0.06	0.73	0.07	0.49	0.01	0.58	0.09
3	0.79	0.10	0.70	0.17	0.50	0.06	0.66	0.06	0.48	0.04	0.49	0.03
4	0.69	0.12	0.53	0.06	0.45		0.60	0.06	0.44	0.02	0.46	0.01
5	0.56	0.02	0.47	0.07			0.55		0.42	0.03	0.45	0.04
6	0.54	0.04	0.41	0.05					0.39	0.04	0.41	0.06
7	0.50	0.06	0.36						0.35		0.35	
8	0.44											

Note. Reading (4) = grade 4 Reading Comprehension, Reading (7) = grade 7 Reading Comprehension, Maps (4) = grade 4 Maps and Diagrams, Maps (7) = grade 7 Maps and Diagrams, Voc (8) = grade 8 Vocabulary, Simulation = simulated data, Eig = eigenvalue, and Diff = difference between consecutive eigenvalues.

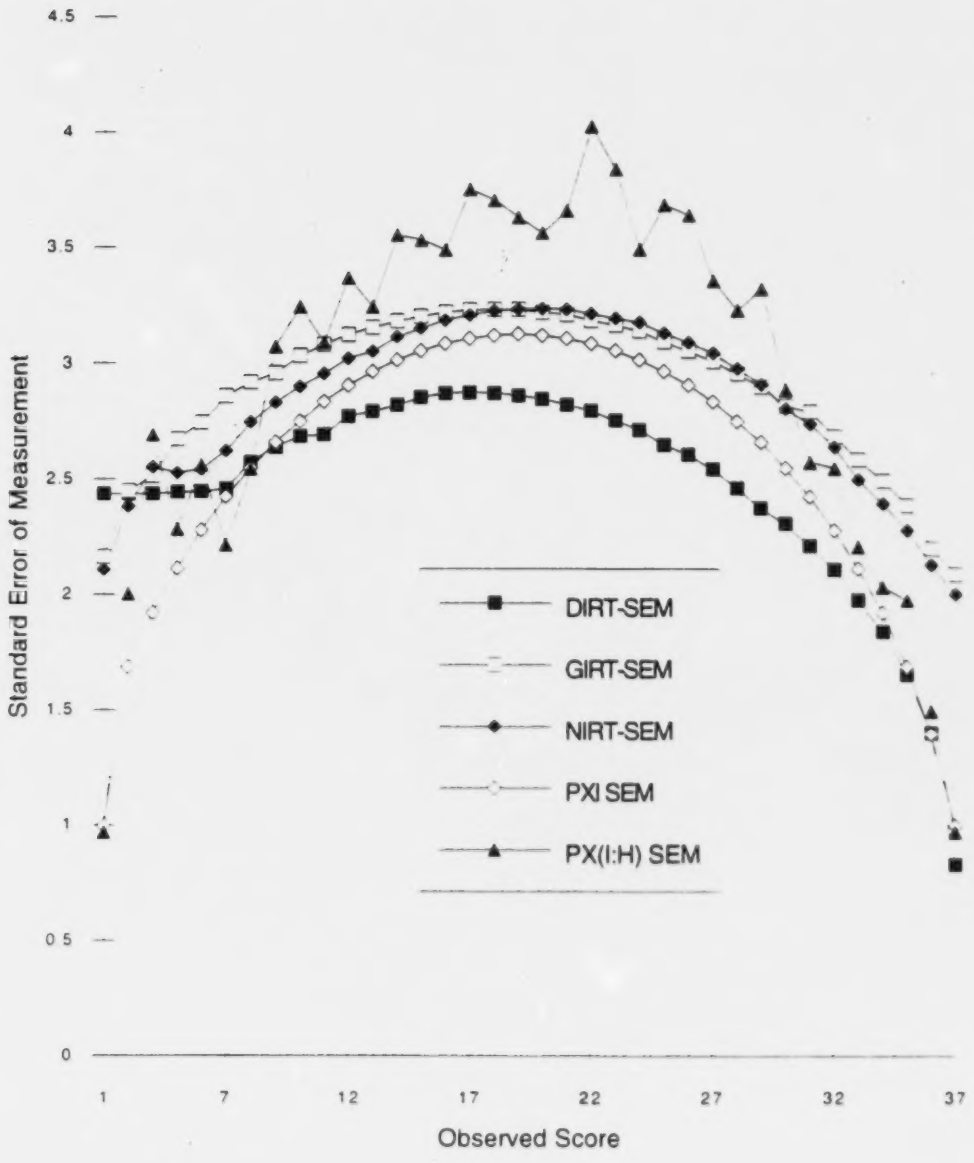
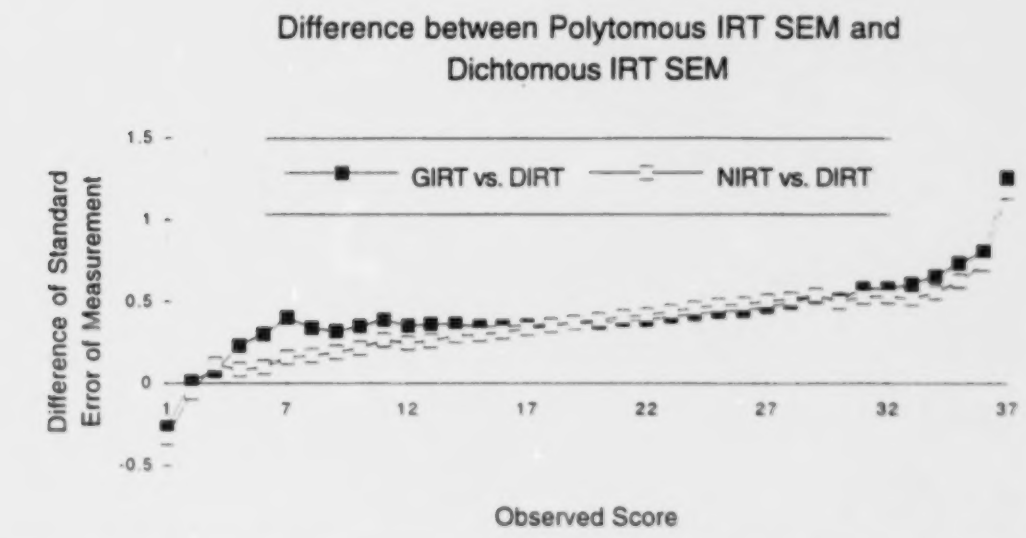


Figure 1. Conditonal standard error of measurement for the Reading Comprehension test (Grade 4) using five estimation methods.



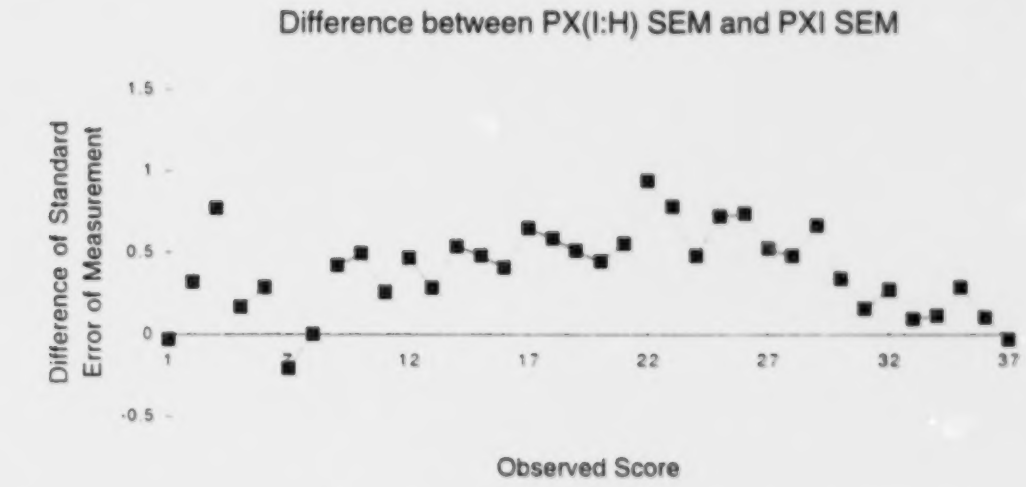


Figure 2. Differences in conditional standard errors of measurement between item-based and testlet-based estimation methods for the grade 4 Reading Comprehension test.

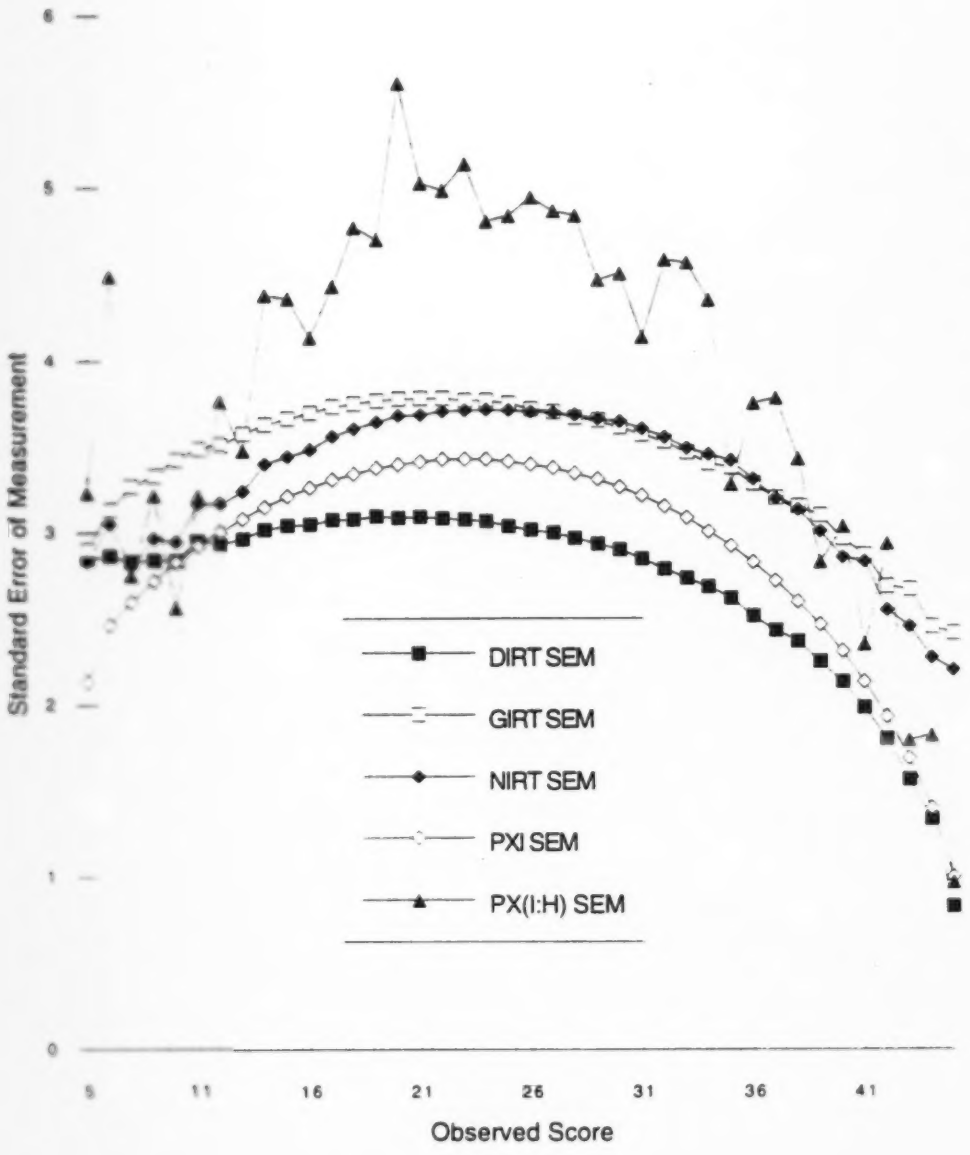
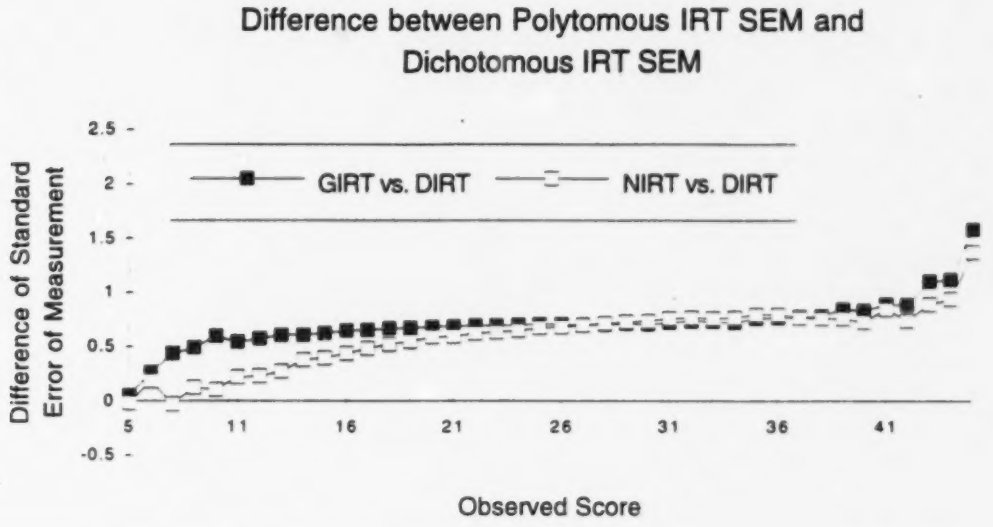


Figure 3. Conditional standard error of measurement for the Reading Comprehension test (Grade 7) using five estimation methods.



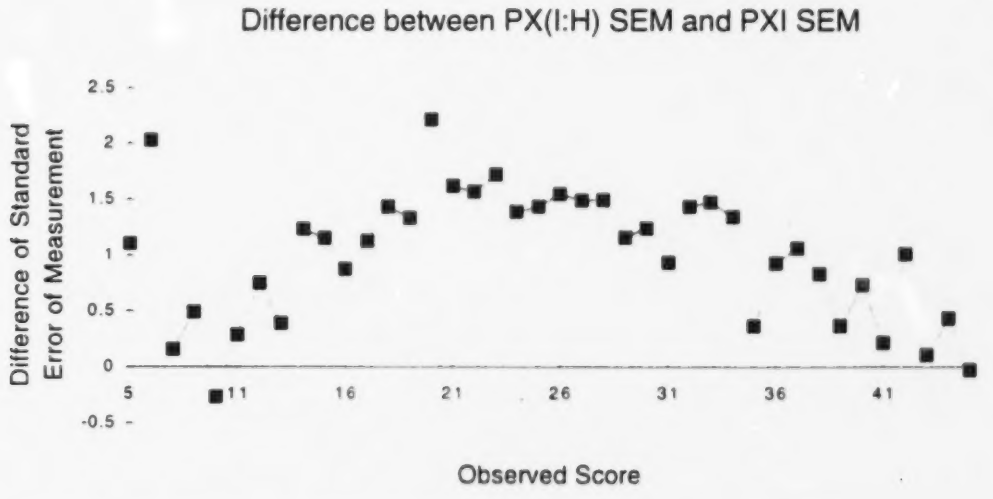


Figure 4. Differences in condtional standard errors of measurement between item-based and testlet-based estimation methods for the grade 7 Reading Comprehension test.

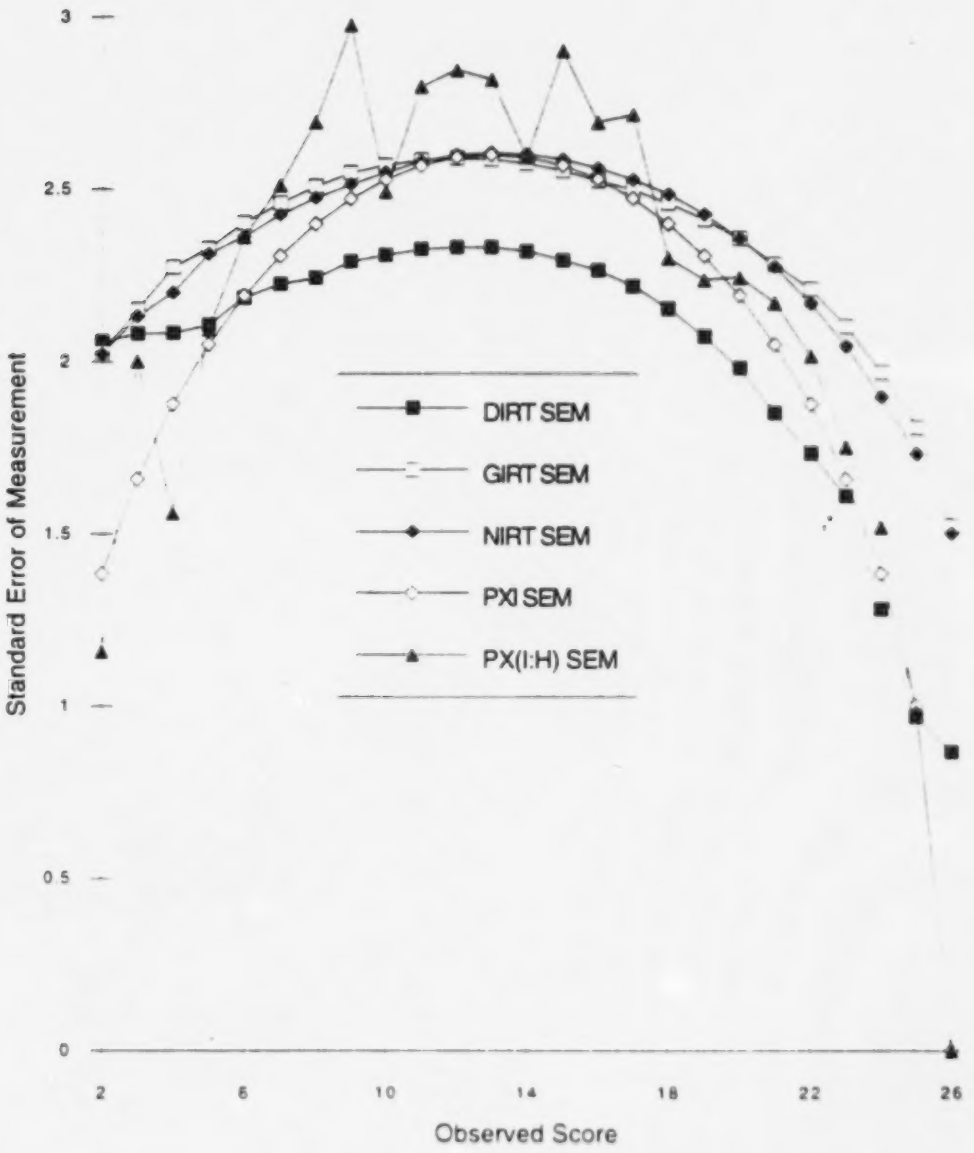
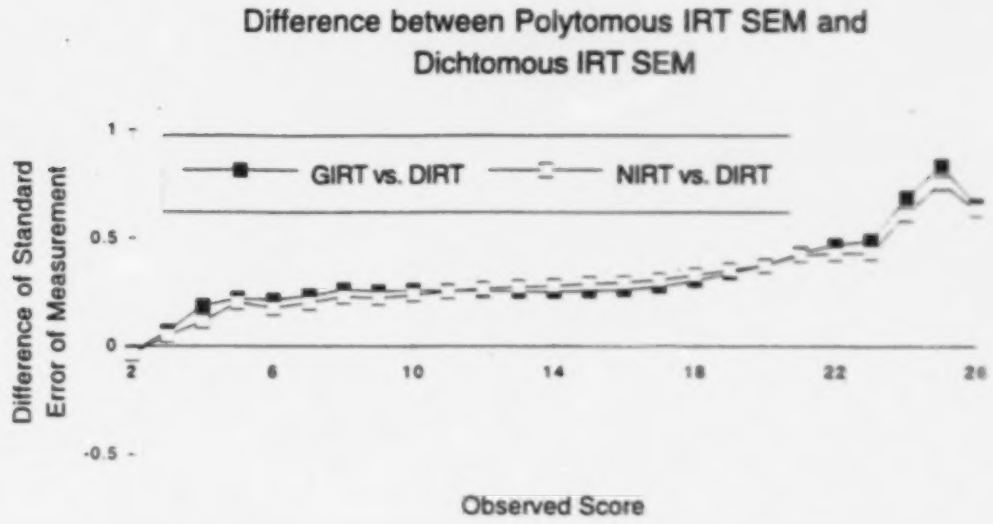


Figure 5. Conditional standard error of measurement for the Maps and Diagrams test (Grade 4) using five estimation methods.



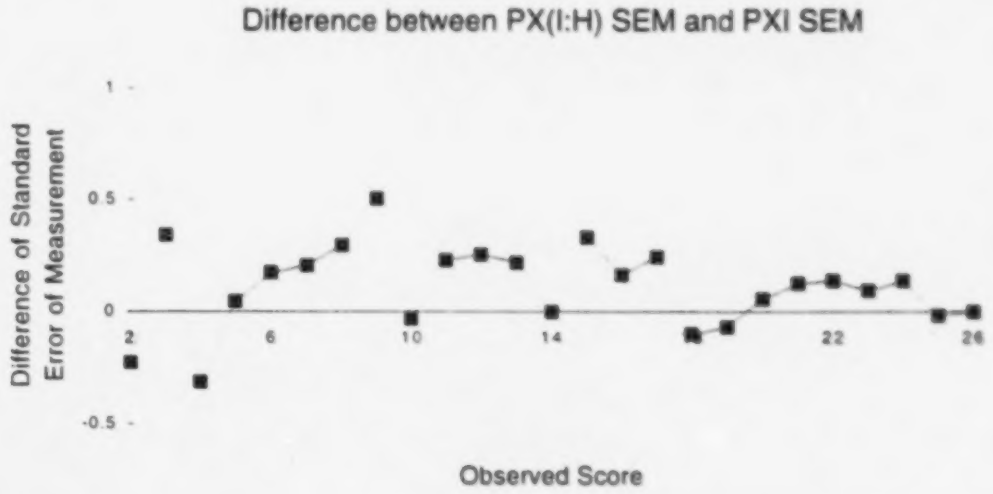


Figure 6. Differences in conditional standard errors of measurement between item-based and testlet-based estimation methods for the grade 4 Maps and Diagrams test.

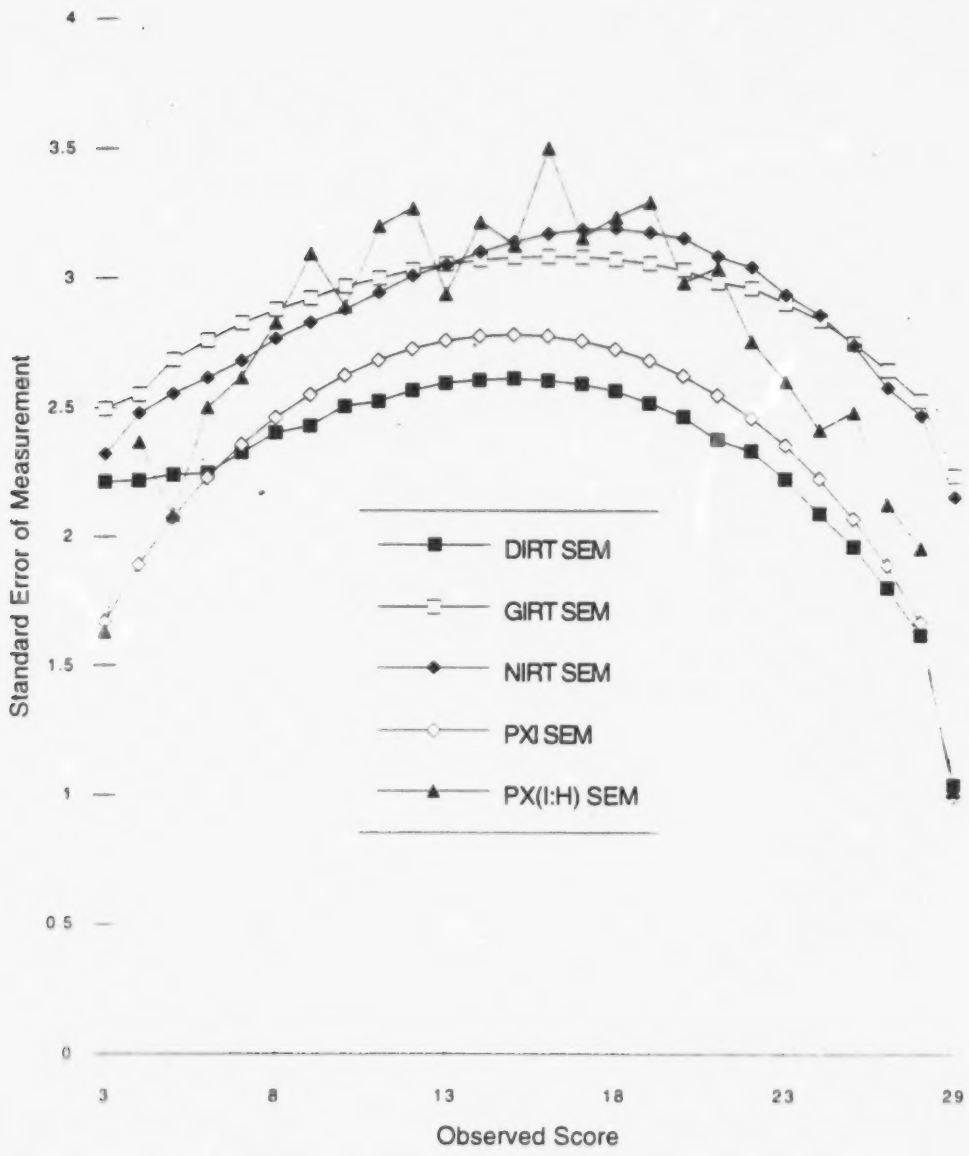
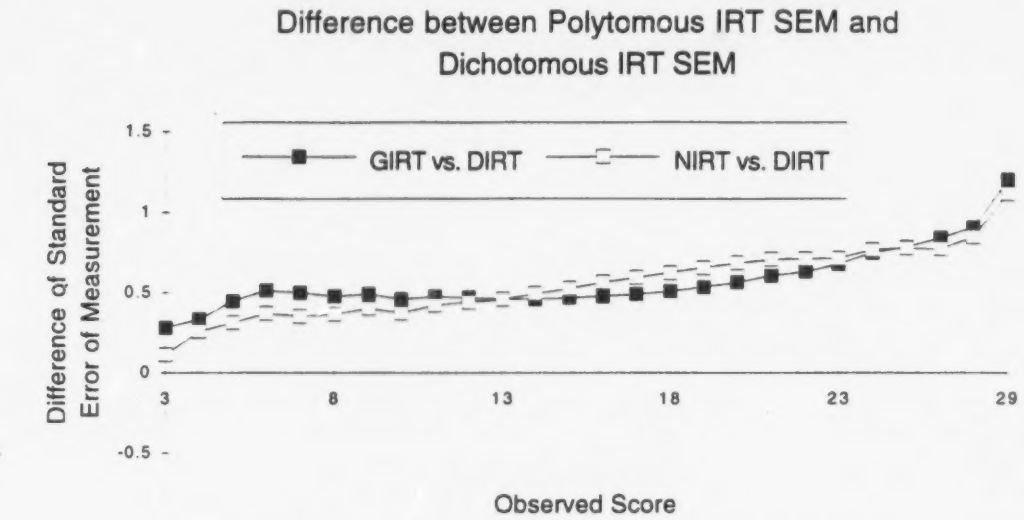


Figure 7. Conditional standard error of measurement for the Maps and Diagrams test (Grade 7) using five estimation methods.



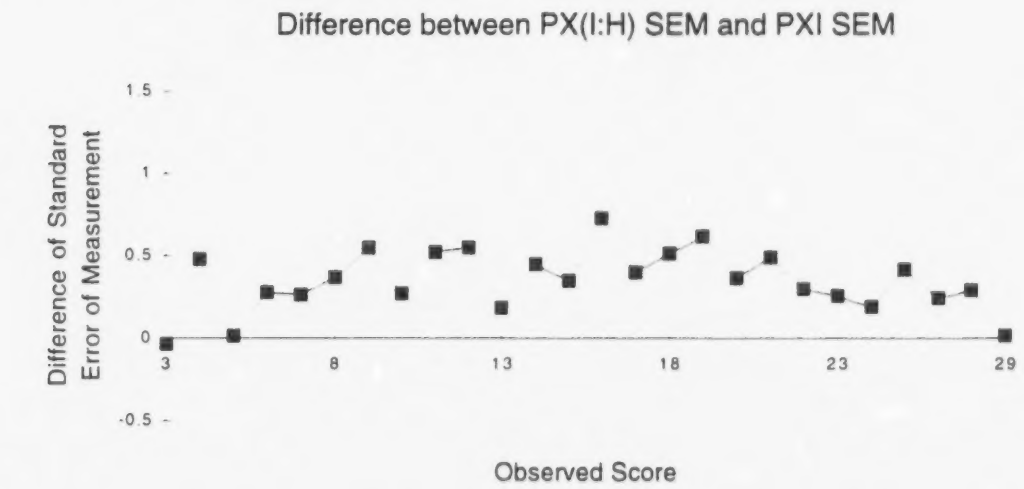


Figure 8. Differences in conditional standard errors of measurement between item-based and testlet-based estimation methods for the grade 7 Maps and Diagrams test.

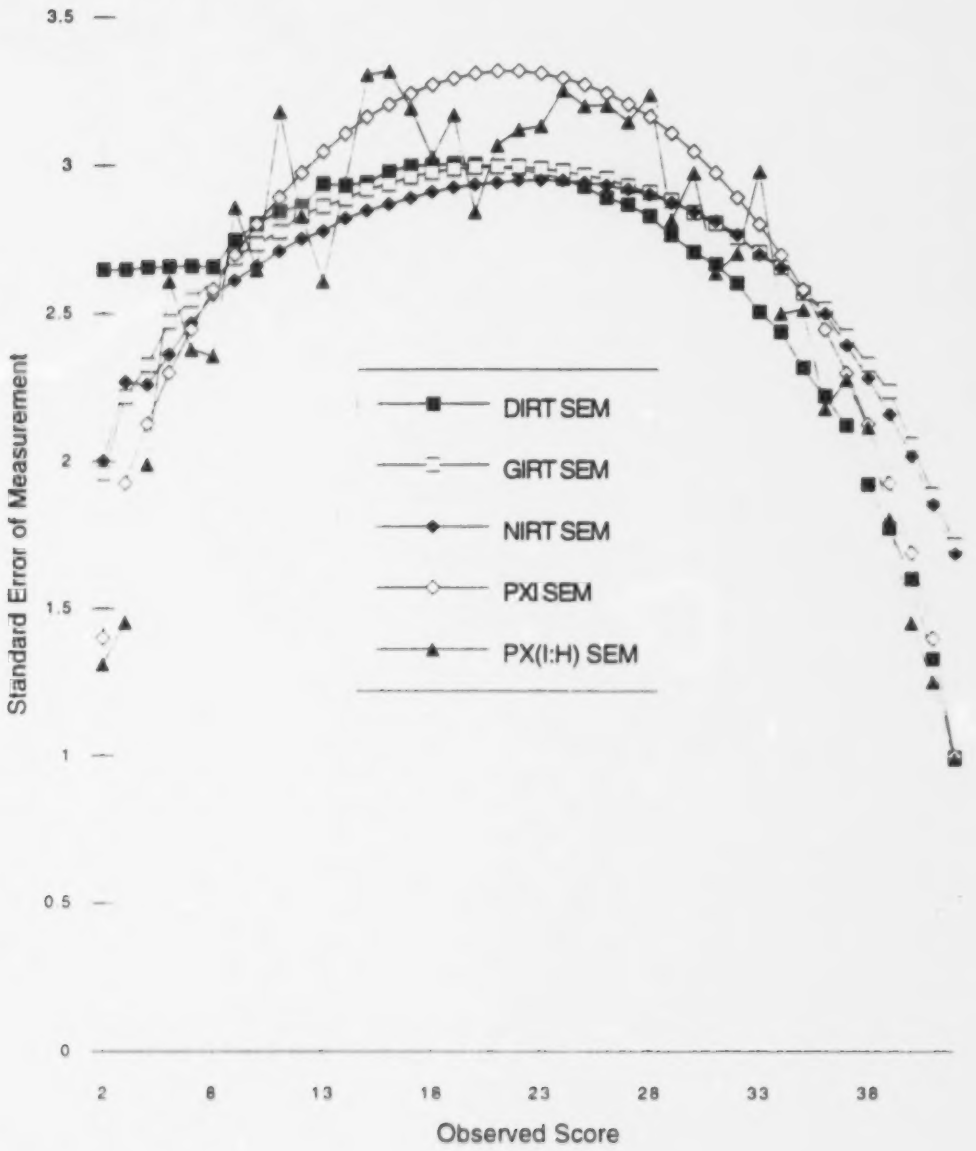
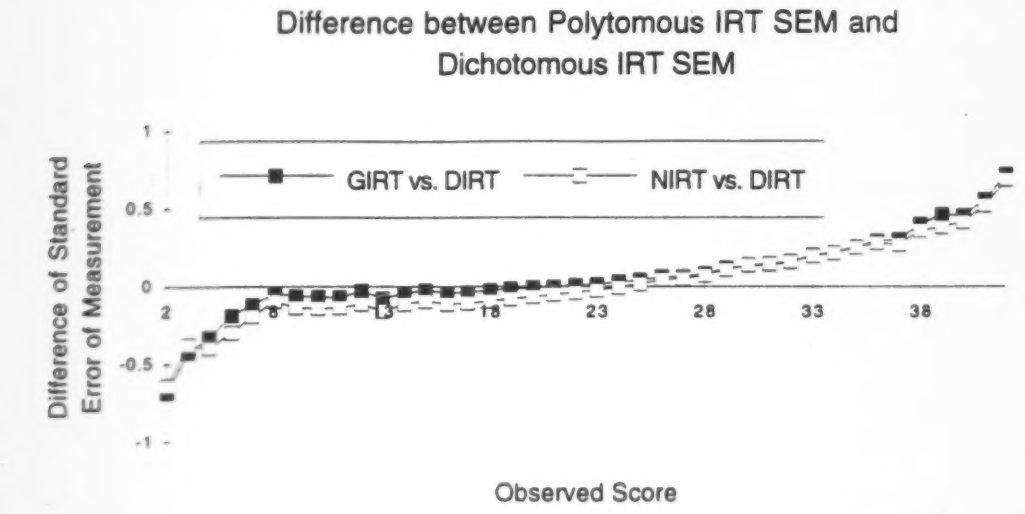


Figure 9. Conditional standard error of measurement for the Vocabulary test (Grade 8) using five estimation methods.



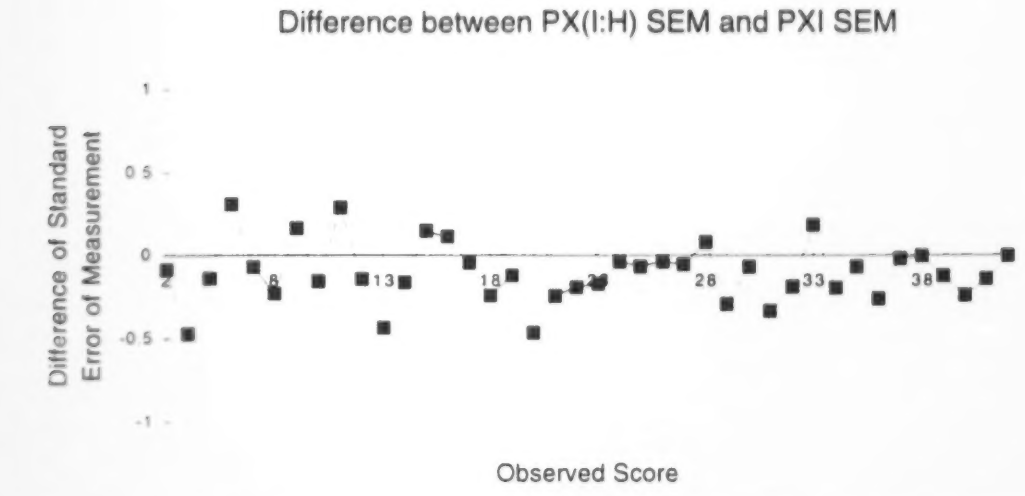


Figure 10. Differences in conditional standard errors of measurement between item-based and testlet-based estimation methods for the grade 8 Vocabulary test.

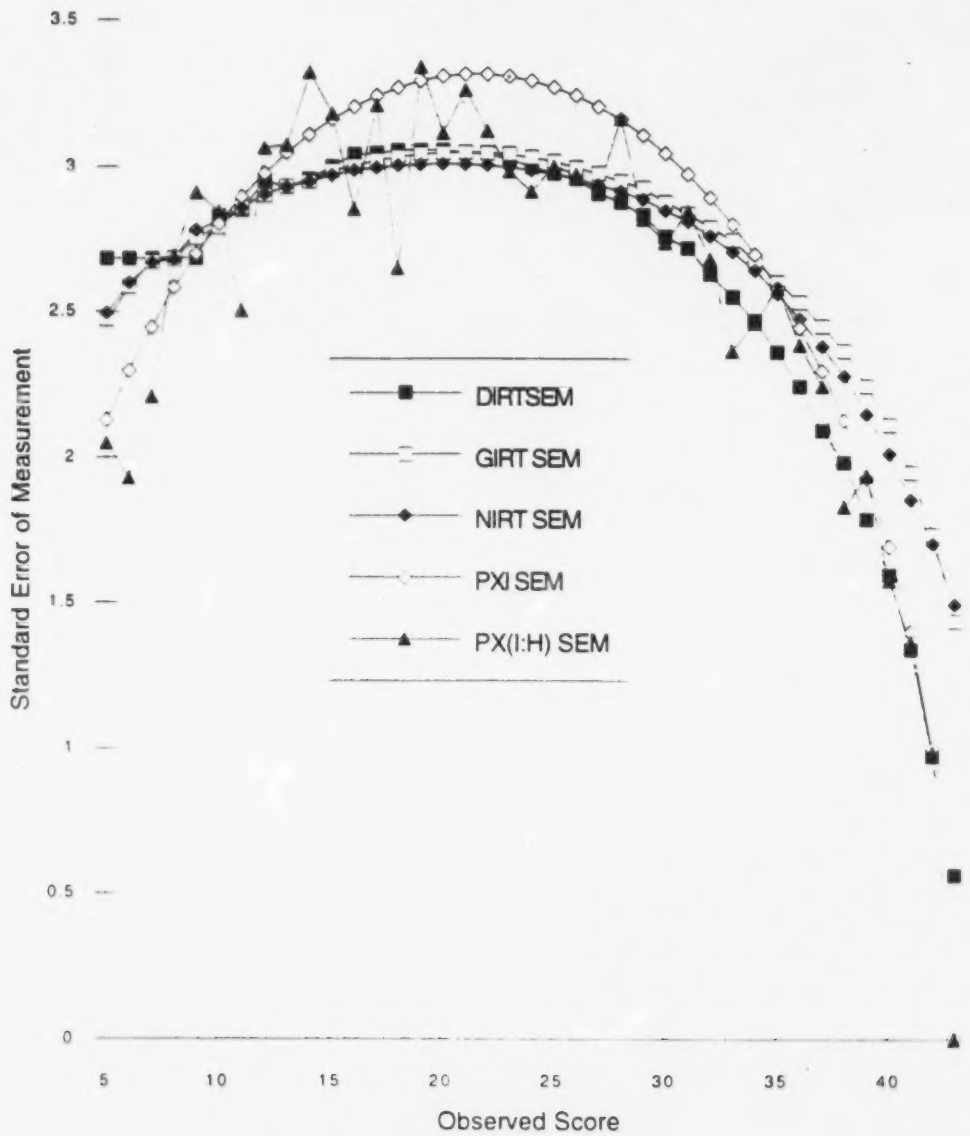


Figure 11. Conditional stnadard error of measurement for the simulated data set using five estimation methods.

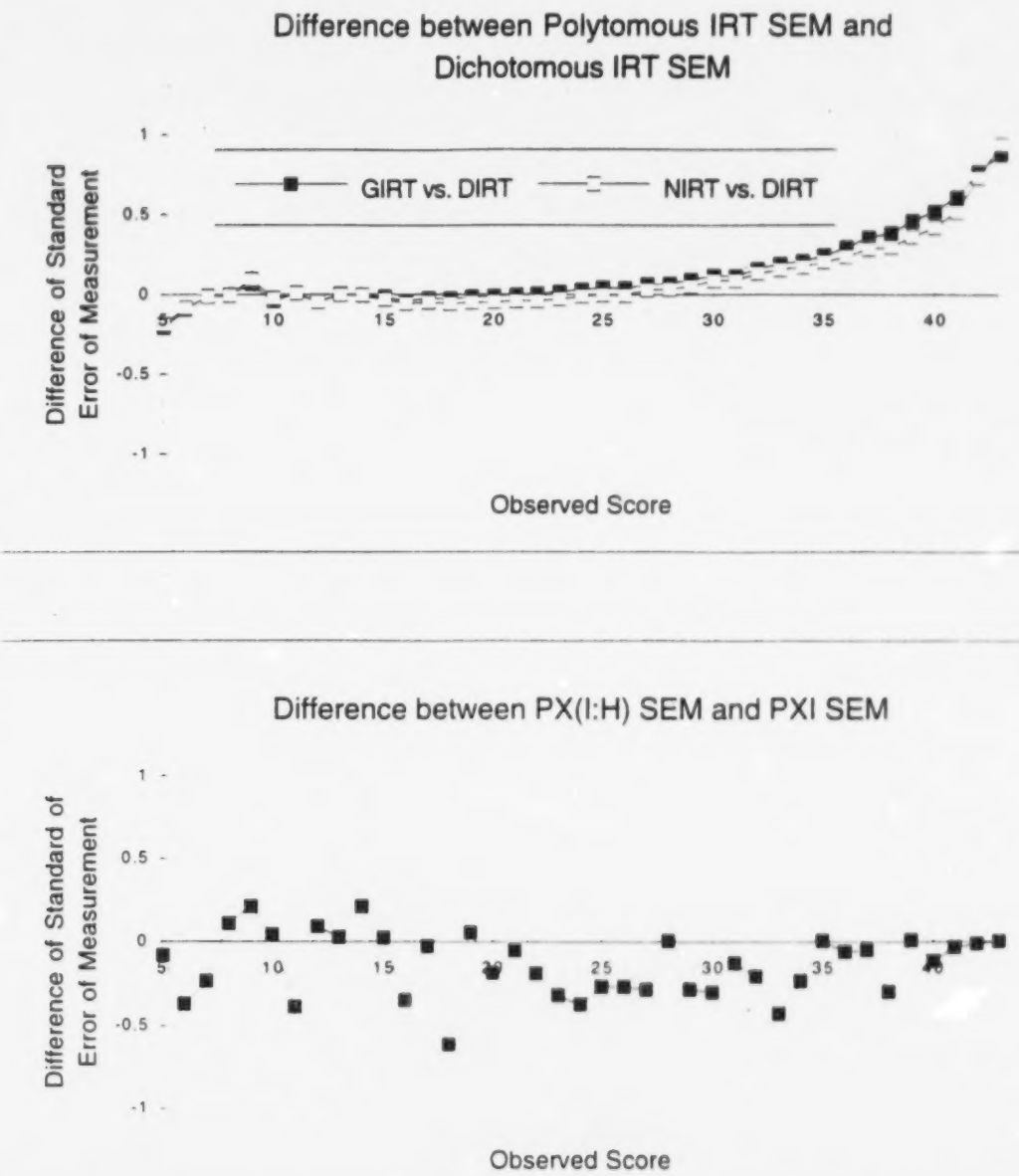

Difference between Polytomous IRT SEM and Dichotomous IRT SEM
GIRT vs. DIRT
NIRT vs. DIRT
Difference of Standard Error of Measurement
Observed Score
Difference between PX(I:H) SEM and PXI SEM
Difference of Standard of Error of Measurement
Observed Score

Figure 12. Differences in conditional standard errors of measurement between item-based and testlet-based estimation methods for the simulated data set.